Biweekly Report 1 and 1.1

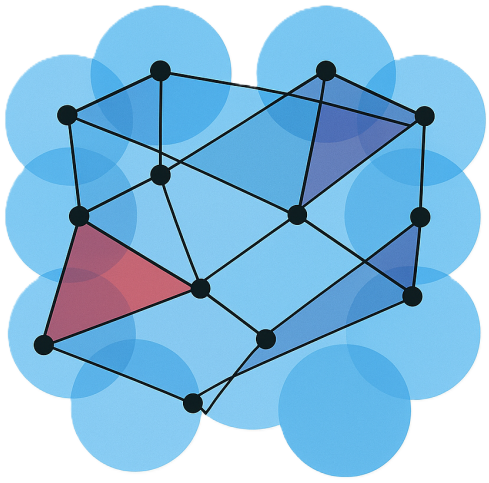
Using Persistent Homology to Compare Pathways and Discussion upon Creating New Species on the Local Machine

Raunak Narwal
Department of Mathematical Sciences
Indian Institute of Science Education and Research, Mohali, 130406, Punjab

August 7 to August 21, 2025

Persistent Homology

Complex structures like chemical reaction networks can be compared by persistent homology by capturing their topological features. A point cloud or network is given as input and we get a barcode or persistence diagram that shows which features persist across which scales as output. To apply persistent homology, we must convert chemical network into a mathematical object. We can compare two chemical networks by comparing their persistence diagrams using bottleneck distance and wasserstein distance.



Comparing Chemical Networks Using Persistent Homology

Chemical reaction networks can be represented as graphs, where metabolites (A substance made or used when the body breaks down food, drugs or chemicals, or its own tissue (for example, fat or muscle tissue)) form the nodes and enzymatic transformations form the edges. Traditional approaces to compare chemical networks focus on pathway overlaps or shared enzymes. These methods often miss out on global structural similarities that are critical to understanding functional equivalence between different biochemical systems. Persistent homology offers a way to adress this limitation by comparing the shape of chemical networks across multiple scales.

In this report, we are comparing two non trivial metabolic networks from KEGG databse using persistent homology. **Phenylalanine, tyrosine and tryptophan biosynthesis (KEGG map00400)** is an aromatic amino acid pathway present in bacteria, fungi and plants. **Terpenoid backbone biosynthesis (KEGG map00900)** is a pathway leading to precursors of steriods and carotenoids which is crucial for membrance stability and adaptation to stress.

These two pathways appear distinct at the biochemical level, their network topologies can be interrogated using persistent homology to reveal that whether they share organization such as recurring loops, feedback structures and modularity. By constructing Vietoris–Rips complexes from metabolite interaction graphs and comparing persistence diagrams using bottleneck and wasserstein distances, our aim is to determine whether these two pathways exihibit topological similatrity that could suggest analogous functional robustness.

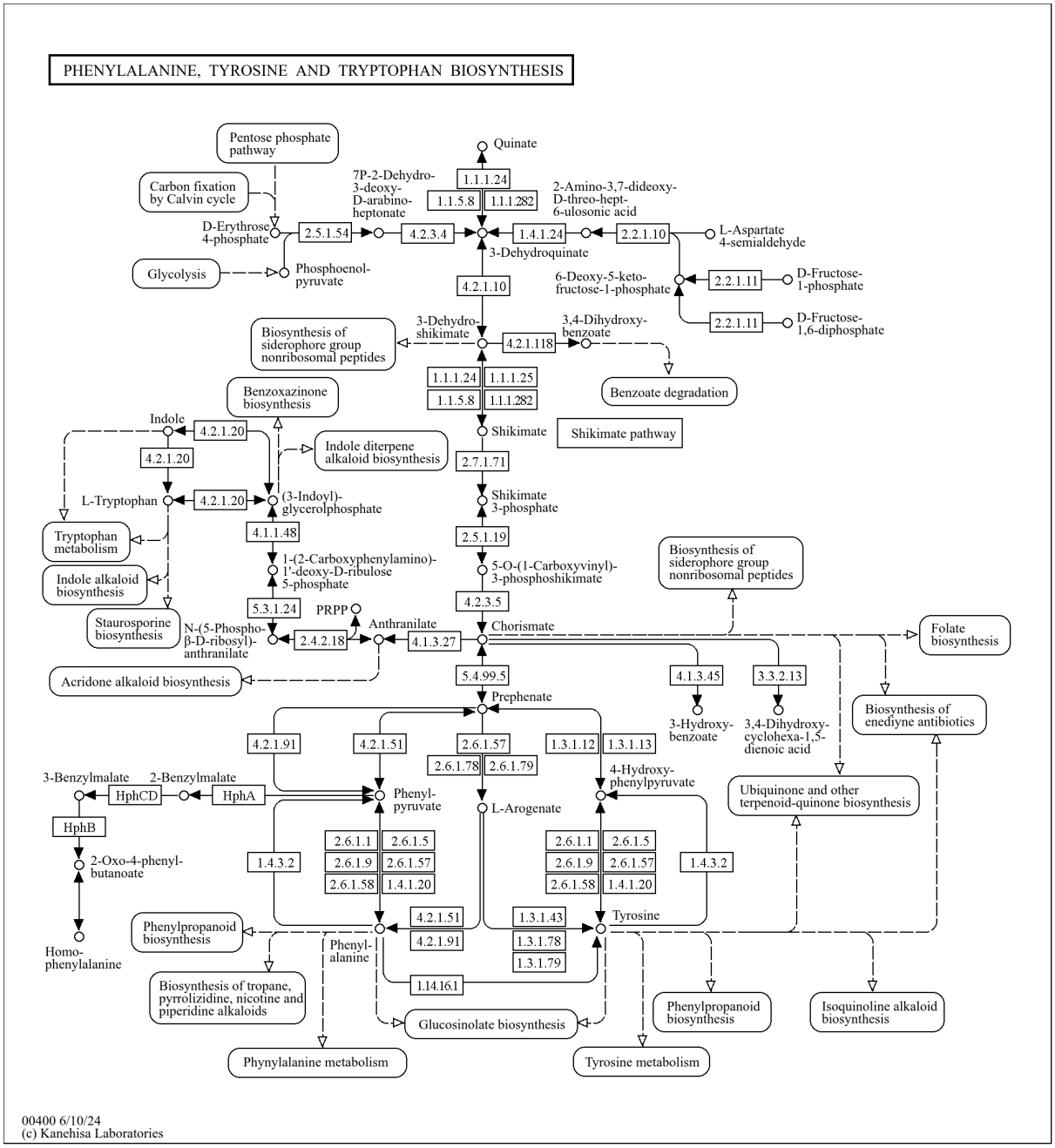


Figure 1: KEGG map00400

To compare the Phenylalanine, tyrosine and tryptophan biosynthesis pathway (KEGG map00400) with the Terpenoid backbone biosynthesis pathway (KEGG map00900), we used a computational pipeline based on persistent homology. Each pathaway is represented as a graph (network), the nodes of the graph are chemical compounds and the edges represent biochemical reactions connecting one compound to another. To apply topological methods, we require to have a notion of distance between compounds in the networks, we calculate the shortest path distances between all pairs of nodes, making a distance matrix. Using the distance matrix, we construct a Vietoris–Rips complex, which is a higher dimensional structure (intuitively) that encodes the connectivity of the network at different distance thresholds. As the threshold is increased, new connections and loops start to form in the complex. With the Rips complex, we compute persistent homology, which

tracks how topological features appear and disappear across scales. H_0 features represent connected components (clusters of compounds) and H_1 features represent loops or cycles in the reaction network. The results are visualized in a persistence diagram, where each point is a topological feature. To compare these pathways, we use distance metrics such as Wasserstein distance (measures the overall distributional difference between features in the two diagrams). If these distances are small, the pathways are similar; if large then they differ significantly. Bottleneck distance measures the largest difference between matched features in the two diagrams.

Demo run: Validating the Persistent Homology Pipeline

Before applying our method to chemical networks, we first validated the pipeline using a demo run. We were unable to obtain KGML files for the two above-mentioned pathways (KEGG Markup Language format). We thought to test our framework on synthethic graphs instead. This ensured that the implementation and computational steps were working as expected. In the demo run, we constructed two simple networks, `Cycle5`: a cycle graph with 5 nodes, which contains a clear topological loop and `Path5`: a path graph with 5 nodes, which is acyclic and therefore lacks loops.

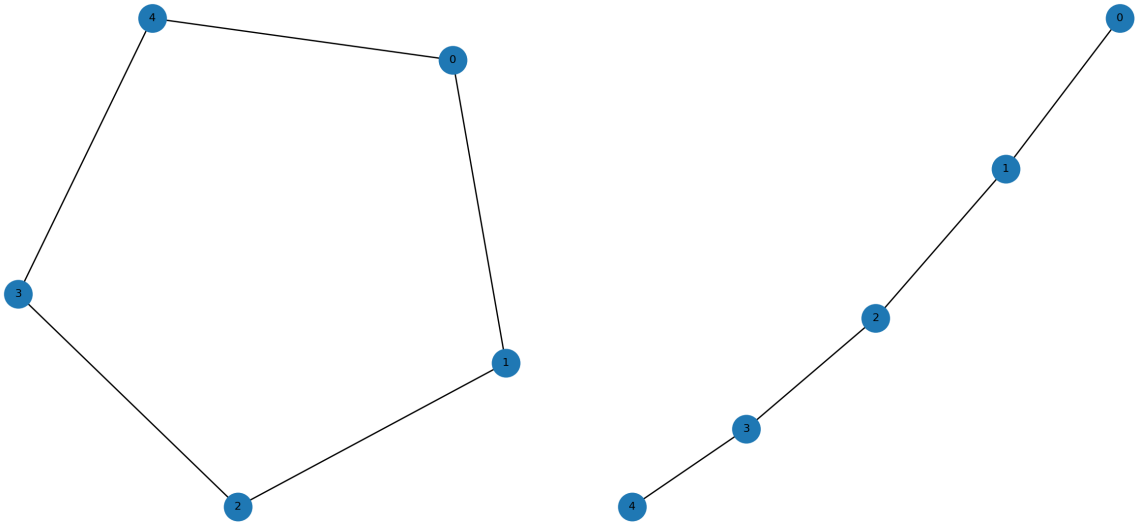


Figure 2: Left: Cycle5 (looped). Right: Path5 (acyclic).

Running the pipeline, we observed that the persistence diagram for Cycle5 clearly shows a non trivial H_1 (corresponding to the loop structure) and the persistence diagram for Path5 contains only H_0 features, reflecting connected components but no higher dimensional holes.

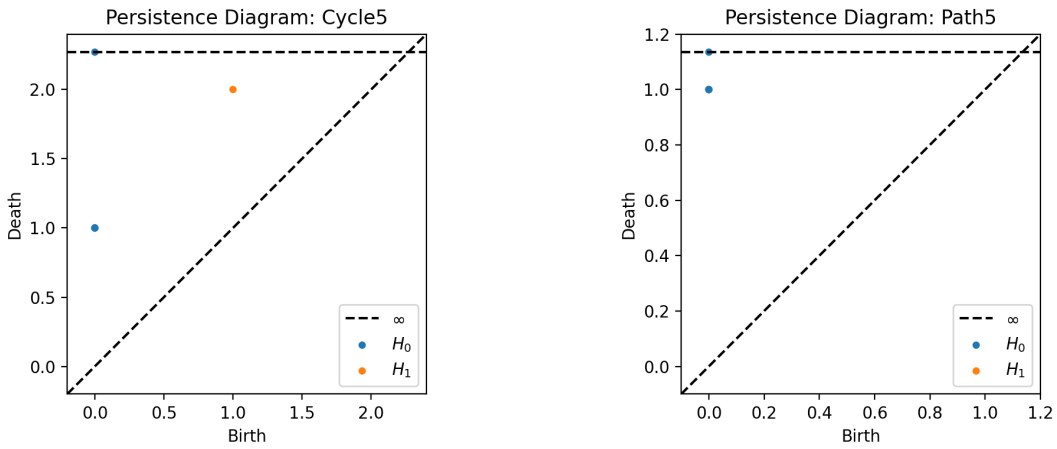
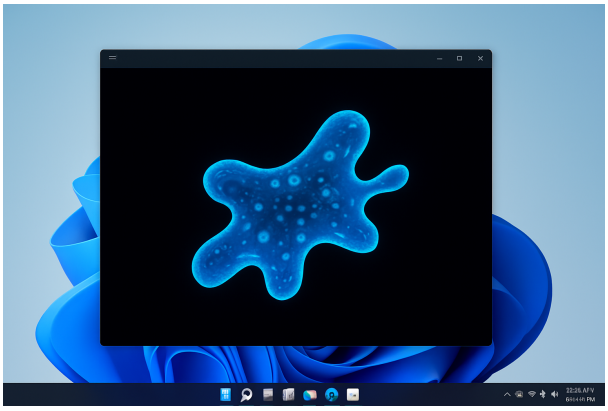


Figure 3: Persistence diagrams from demo run.

How Prokaryotic Cells work without a Nucleus

Prokaryotic cells do not contain a nucleus or membrane bound organelles, still they manage to sense their environment, process information and adapt to change in conditions. Their genetic material floats freely in the cytoplasm, still they function collectively as a from of distributed computation. Our interpretation is that these cells are just clump of chemicals, so why not we simulate them on the computer, why not we design our own similar like cells or newer species? Gene expression is coupled directly to translation, creating rapid response to environmental stimuli. Feedback looks in pathways allow these cells to learn from experience, adjusting their internal state based on past exposure to stress and toxins. This behaviour contrasts with eukaryotic cells, where nucleus actts as the control hub of the cell. Instead, prokaryotes rely on the emergent behavior of thousands of interacting molecules.



Creating virtual life on the laptop

Getting inspired by prokaryotic adaptability, we ask this question, whether simplified digital organism can be designed on a computer using rule based chemical interaction networks. These speciies could consist of virtual molecules that follow reaction rules, forming pathways. Environment constraints can be assigned such as temperature shifts, scarcity of nutrients and stress. We can observe how these virtual organisms adapt over

time. Persistent Homology can play a role here serving as a tool to track how topological structures of internal chemical network changes over time.
a very exciting application of this idea lies in explosring life in extraterrestrial environments, such as Mars and Moon. Condition at Mars are characterized by low temperatures, high radiation , limited liquid water and a predominantly carbon dioxide rich atmosphere. these variables could be encoded as environmental constraints, we can evolve chemical networks that survive in such conditions. If a clump of chemicals figured out how to survive in a nitrogen and oxygen rich atmosphere, it would be foolish to think that martian conditions could not host animals, yes they could not host earth like organisms, but we see a strong posibility that there must be some other sort of CRNs (chemical reaction networks) that make martian survival feasible. **Nature thrives in harsh conditions; nature is metal**. For instance, instead of oxygen dependent energy pathways, a martian adapted digital organism might evolve networks favoring methanogenensis or perchlorate reduction. This discussion would be continued in the next biweely report.

Agendas for next report

1. Comparing chemical reaction networks from KEGG database
2. Figuring out if we can call KEGG .kmgl files from local machine and simulate permutations of comparisons, we might face constraints like kegg's bulk distribution policy.
3. Comparison of dynamic similarity between chemical reaction networks.

Biweekly Report 1.1

We are making this report to compensate for the absense during mid semester examination from 1 September to 4 September.

Event	Timeline
Biweekly Report 1	21 August
Biweekly Report 1.1	25 August
Mid semester exams	1-3 September
Biweekly Report 2	~~3 September~~ 7 September

Comparing real chemical networks

Using persistent homology on the shortest parth distance matrices of the two KEGG pathways, Phenylalanine, tyrosine and tryptophan biosynthesis (KEGG map00400) and Terpenoid backbone biosynthesis (KEGG map00900) , we observed that the two pathways are close but not identical in their topological signatures. Both of the networks have similar global size (Pathway1 : 98 nodes and 355 edges and Pathway2 : 106 nodes and 275 edges) and comparable graph diameters (13 and 12 respectively). The bottleneck distance for both H_0 and H_1 is 2.0 (normalized to 0.154 when divided by the larger diameter 13) , which indicates only modest shifts in the largest persistence features between the diagrams. Below is the summary JSON file.

```
{
  "GraphA": {
    "name": "Pathway1",
    "nodes": 98,
    "edges": 335,
    "diameter": 13,
    "graph_plot": "outputs\\Pathway1_graph.png",
    "diagram_plot": "outputs\\Pathway1_tda_diagram.png"
  },
  "GraphB": {
    "name": "Pathway2",
    "nodes": 106,
    "edges": 275,
    "diameter": 12,
    "graph_plot": "outputs\\Pathway2_graph.png",
    "diagram_plot": "outputs\\Pathway2_tda_diagram.png"
  },
  "Metrics": {
    "H0_bottleneck": 2.0,
    "H0_bottleneck_norm": 0.15384615384615385,
    "H0_wasserstein": Infinity,
    "H0_wasserstein_norm": Infinity,
    "H1_bottleneck": 2.0,
    "H1_bottleneck_norm": 0.15384615384615385,
    "H1_wasserstein": 2.8722813232690143,
    "H1_wasserstein_norm": 0.22094471717453956
  }
}
```


What is H_0 and what do the results tell?

In persistent homology, the 0-dimensional homology group, or H_0, tracks the number of connected nodes or clusters in a dataset. It provides a count of these clusters by observing how they evolve as a proximity parameter or scale. Initially each data point is its own separate component. As the scale grows nearby nodes connect causing them to merge. The "birth" of a component is when it first appears at the start of 0-dimension and its "death" occurs when it mergers into another component. The most significant , robust clusters are those that persist for a longer time.

The bottleneck distance of 2 (norm = 0.154) shows that the most persistent connected component feature moves by about two units between two networks. The reported wasserstein distance is infinity in summay, this is because one of the diagrams contains some points that persist to infinity. This tells that there is a structural difference in how components appear between two pathway graphs.

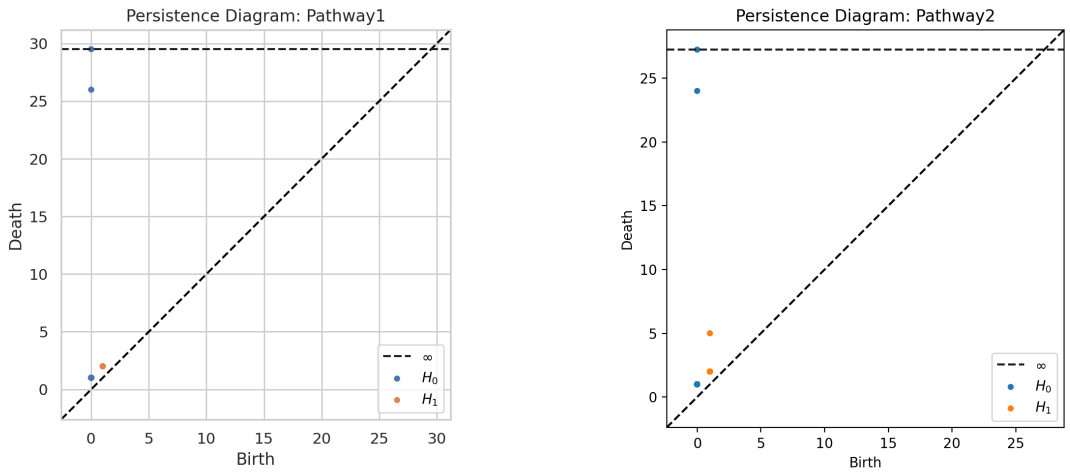


Figure 4: pathways tda diagram

What is H_1 and the results for H_1

H_1 in persistent homology, is the 1 dimensional homology group. It tracks the presence of loops or 1-dim holes within a dataset as the scale parameter changes. As the system evolves, H_1 records when such loops appear ("Birth") and when they are filled in or become boundaries ("Death"). The barcodes or persistence diagrams for H_1 highlight significant topological features like tunnels in 3D data or holes in the 2D data. Persistent H_1 provides insight into the multi scale structure.
The bottleneck distance is 2 and the wasserstein distance is 2.872 (norm = 0.221). Th finite wasserstein indicates overall distribution of loop persistences has changed, some cycles in one pathway appear earlier or dier later relative to the other, but there is not a big mismatch. The normalized wasserstein means the typical change in cycle persistence is about 22 percent of the larger graph diameter (greatest distance between any pair of vertices of a connected graph), it is useful to interpret effect size relative to nework scale.

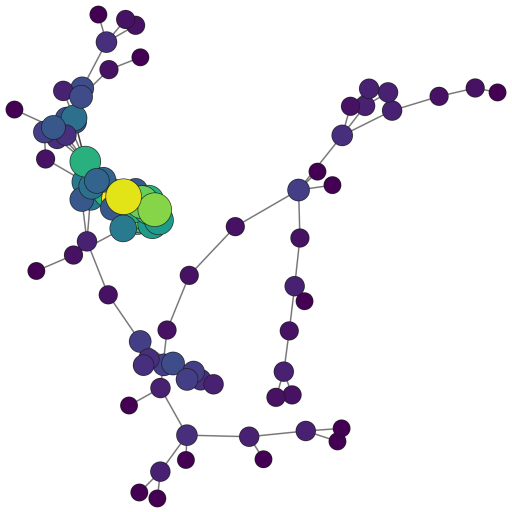


Figure 5: map00400 : Phenylalanine, tyrosine and tryptophan biosynthesis graph

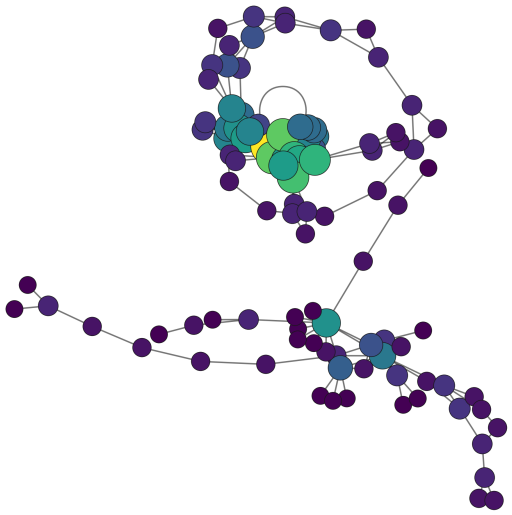


Figure 6: Terpenoid backbone biosynthesis (KEGG map00900)

Methodology Notes

the pipeline accepts either CSV or KGML pathway files. When a KGML is supplied, the script extracts *entry* elements into node names, parses *relation* elements as pairwise edges, and expands *reaction* elements into edges from each substrate to each product. It writes a three column CSV $(sources, tagret, type)$ that preserves whether an edge came from a relation or reaction. We have a small validation function that compares nodes and edges counts and overlaps that with the original KGML graph and the CSV derived graph to ensure conversion preserved topology. Both CSV and KGML derived graphs are loaded into $NetworkX$ as undirected graphs with edge weights. The pipeline converts directed KGML into an undirected represntaiton for the topological analysis, the script then computes all pairs shortest path matrix that fills graph disconnected pairs with a large finite value so that ripser can accept a full finite distance matrix. Then, persistent diagrams are computed from the shortest path distance matrix using ripser, $ripser(..., distance_matrix = True)$ (maximum homology dimension is defaulted to 1). To avoid including extremely long distances in the Vietoris Rips filtration, we automatically set a threshold (95 percentile of finite distance). The code also saves raw *.npy* and $JSON$ of the digrams and also produces the plotted persistence diagrams